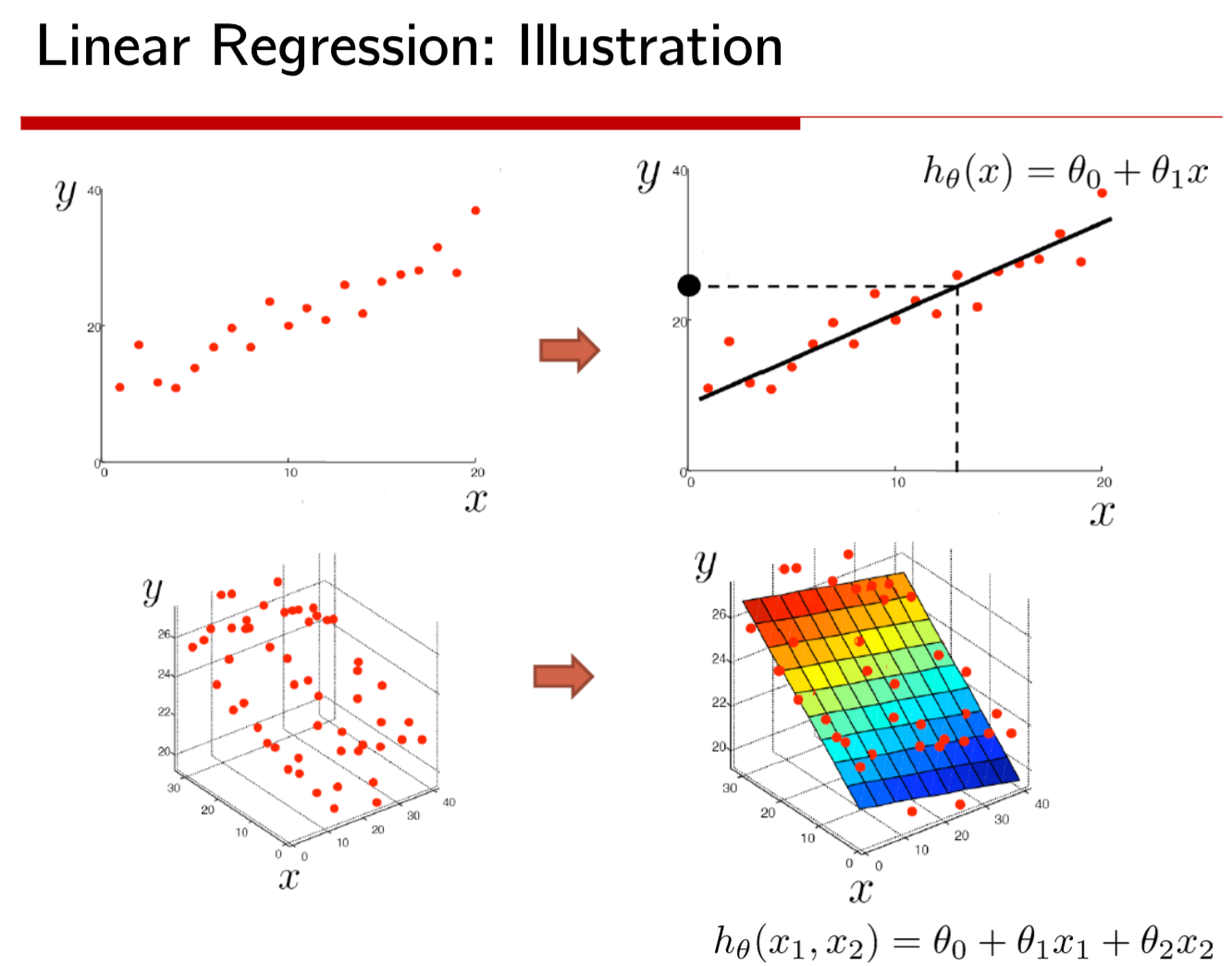
**Home Work 3-Data Mining**

**Apeksha Hada**

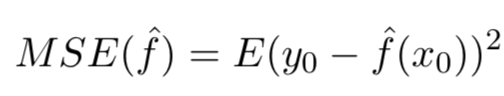
1. **Answer the following? [50 pts]**
2. **Write and explain about the Linear Regression and its equation [3 pts]**

Answer:



1. Explain in detail about the loss function of linear regression, R2, Adjusted R2 used in the Linear Regression and what is the need for Adjusted R2? [12 pts]

Answer: Loss or error or MSE



**R2:** R-squared is a statistical measure of how close the data are to the fitted regression line. R-squared is a goodness-of-fit measure for linear regression models.

**Adjusted R2:** Problems with R2:

1. Plot X vs Y in Scatter plot from data in Table 1 and comment on the relation of X vs Y using Covariance, Correlation. Please comment on Covariance and correlation values [5 Pts]

|  |  |
| --- | --- |
| X No of Weeks | Y Avg Sales |
| 6 | 526 |
| 3 | 421 |
| 6 | 581 |
| 9 | 630 |
| 3 | 412 |
| 9 | 560 |
| 6 | 434 |
| 3 | 443 |
| 9 | 590 |
| 6 | 570 |
| 3 | 346 |
| 9 |  |

Table 1: X(No of Weeks) vs Y(Avg Sales)

<http://scaryscientist.blogspot.com/2015/10/covariance-and-correlation.html>

Answer: Covariance can be classified as positive covariance (two variables tend to vary together) and negative covariance (one variable is above or below the expected value compared to another variable)

d) Perform Linear regression on the following data using Python? and print β0, β1 values in equation y= β0+ β1\*x. Please write down what is your understanding from those values. [10 Pts]

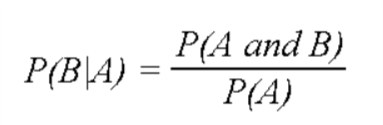
e) What are different evaluation metrics available for predicting the performance of the Linear Regression? Evaluate all those methods on the given dataset in Table 1 and also please print out the accuracy, R2, Adjusted R2 [10 pts]

f) Print ANOVA (Analysis of Variance) table and Parameter Estimates for the given data and explain your understanding. [See hints and explanation for what I am looking for] [10 pts]

1. **Answer the following [30 Pts]**

**a) What is Conditional probability, Marginal probability and Joint probability? Write their mathematical formulas and give one example each. [5 pts]**

Answer: The conditional probability of an event B is the probability that the event will occur given the knowledge that an event A has already occurred. This probability is written P(B|A), notation for the probability of B given A.

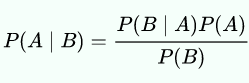


**Marginal probability:** Probability of a single event irrespective of any other event. For instance, the probability of a coin flip giving a head is considered a marginal probability because we aren’t considering any other events.

**Joint probability:** The probability of two events occurring simultaneously. The probability of the intersection of A and B may be written p(A ∩ B). Example:  the probability that a card is a four and red =p(four and red) = 2/52=1/26.  (There are two red fours in a deck of 52, the 4 of hearts and the 4 of diamonds).

**b) Explain what is Baye’s rule with the formula and what is prior, posterior, likelihood and marginal probability in the Baye’s rule. [10 pts]**

Answer: **Bayes’ theorem** or **Bayes’ rule** describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if cancer is related to age, then, using Bayes’ theorem, a person's age can be used to more accurately assess the probability that they have cancer than can be done without knowledge of the person’s age.



P(A|B): Conditional probability: Posterior

P(A) and P(B): Marginal Probability: Prior

P(B|A): Conditional probability: Likelihood.

The posterior is proportional to the prior times the likelihood.

**c) What is Naive Bayes algorithm and how is related or derived or inspired from Bayes rule? [5 pts]**

Answer: **Naive Bayes** is a simple technique for constructing classifiers: models that assign class labels to problem instances, where the class labels are drawn from some finite set.

All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

**Relation with Bayes rule**: The classification is carried out by calculating the posterior probabilities and finding the hypothesis with the highest probability using Bayes rule. It is referred to as naive because it assumes all features are independent, which is rarely the case in real life.

**d) Perform Naive Bayes algorithm on the below dataset in python in which you can classify whether a Red Domestic SUV is stolen or not as shown in 2.2. [10 pts]**

